TMA4267 Linear Statistical Models V2017 (L16)

Part 3: Hypothesis testing and analysis of variance
Multiple testing [note]

Mette Langaas

Department of Mathematical Sciences, NTNU

To be lectured: March 14, 2017

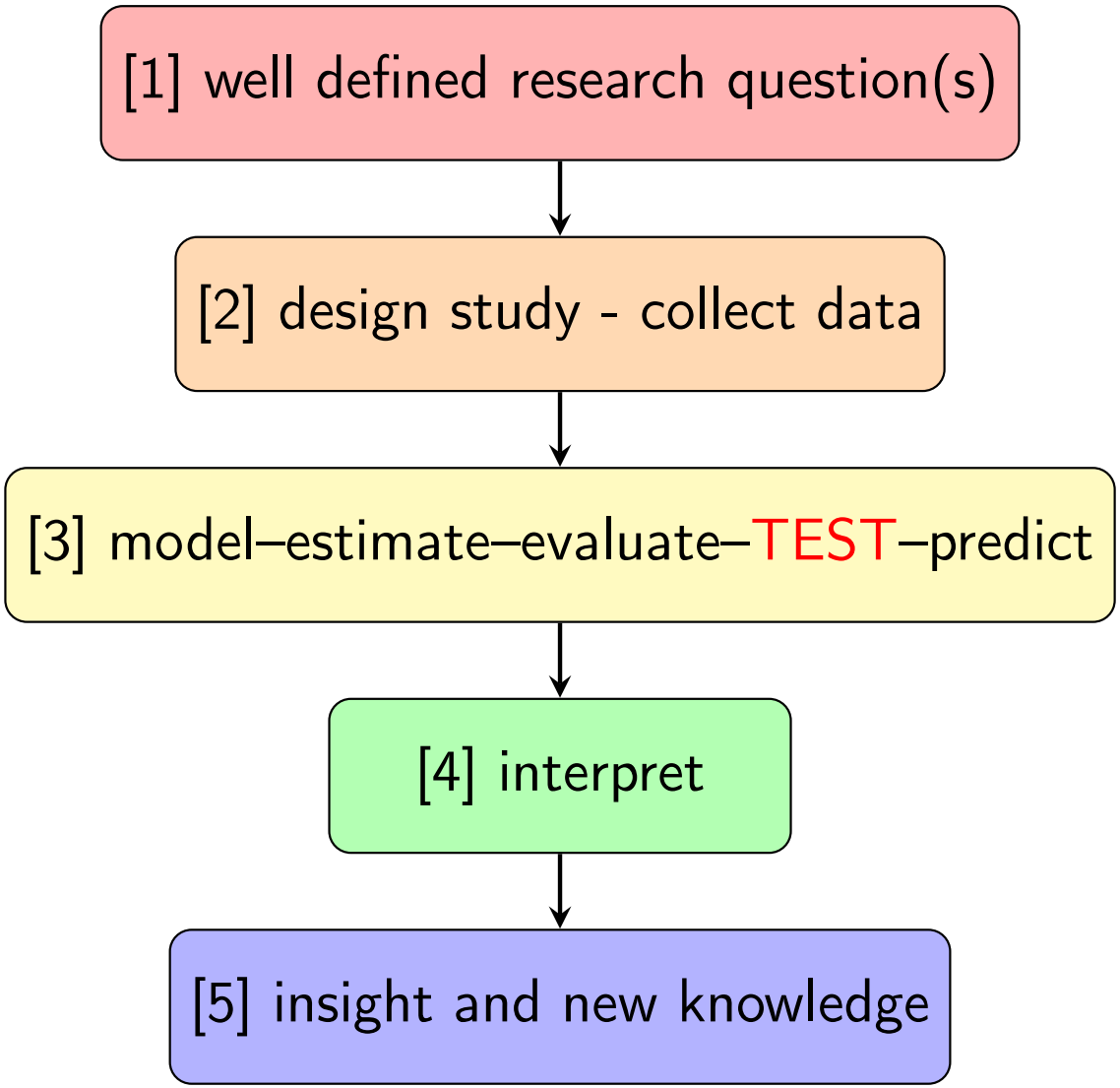
Today: Multiple testing

- Single hypothesis testing: H_0 and H_1, test statistic and p-value.
- Controlling Type I error (false positive findings) by selecting a significance level.
- Properties of p-values from true and false null hypotheses.
- Testing many hypotheses: why?
- Generalizing the type I error from single to multiple hypothesis testing: FWER and FDR.
- Two methods (Bonferroni and Šidák) that control the FWER
- Summarizing Part 3 with a quiz.

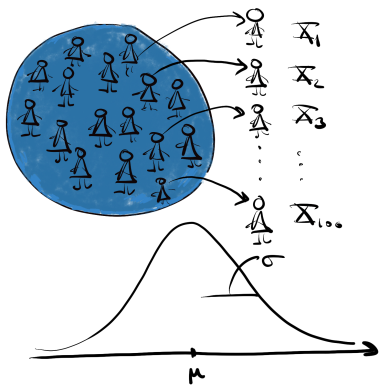
Basal metabolic rate and the FTO-gene

- The gene called FTO is known to be related to obesity
- The basal metabolic rate says how many calories you burn when you rest (hvilemetabolisme).
- Data has been collected for 101 patient from the obesity clinic at St. Olavs Hospital.
- Research question: is there an association between the variant of the FTO gene of the patient and the basal metabolic rate?
- Regression setting, other covariates include age, sex, weight, height, BMI, diet, exercise level, smoking, etc.

The scientific process



Hypothesis testing example (from L13)



- We draw a random sample of size $n = 100$ from the blue population and measure systolic blood pressure: $X_1, X_2, \ldots, X_n$.
- Test statistic: $\bar{X} \sim N(120, 1)$ when H_0 is true.
- We find that $\bar{x} = 122$ mmHg.
- Data: $n = 100$, $\bar{x} = 122$, gives a p-verdi=0.02.

Hypothesis testing example (from L13)

Questions:

- How have I calculated this p-value?
  $P(\bar{X} > 122 \mid H_0 \text{ true})$.
- How can I interpret this p-value?
  Informally, a p-value is the probability under a specified statistical model that a statistical summary of the data would be equal to or more extreme than its observed value.
- Should I conclude that $\mu > 120$?
  Yes, if you choose significance level higher than 0.02. But, you should also report a (two-sided) confidence interval for μ:
  Here $[120.04, 123.96]$.

Single hypothesis testing set-up

	H_0 true	H_0 false
Not reject H_0	Correct	Type II error
Reject H_0	Type I error	Correct

Two types of errors:

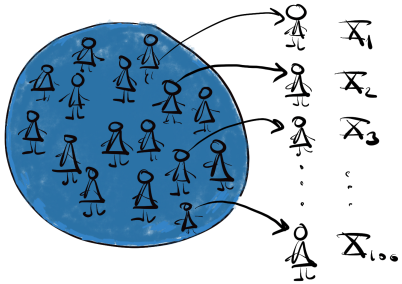
- False positives = type I error =miscarriage of justice.
- False negatives = type II error= guilty criminal go free.

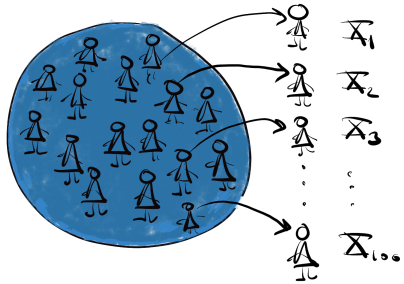
The significance level of the test is α.
We reject the null hypothesis when the *p*-value is *below* α.

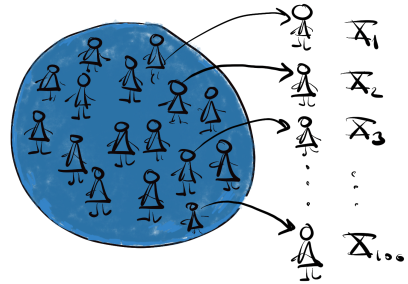
We say that : Type I error is "controlled" at significance level α.

The probability of miscarriage of justice (Type I error) does not exceed α.

Repeating the blood pressure experiment

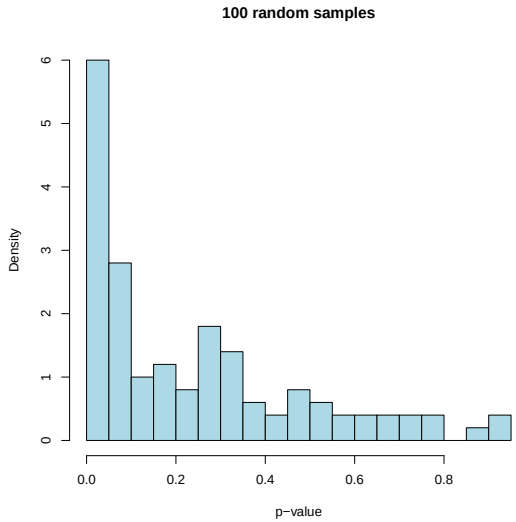


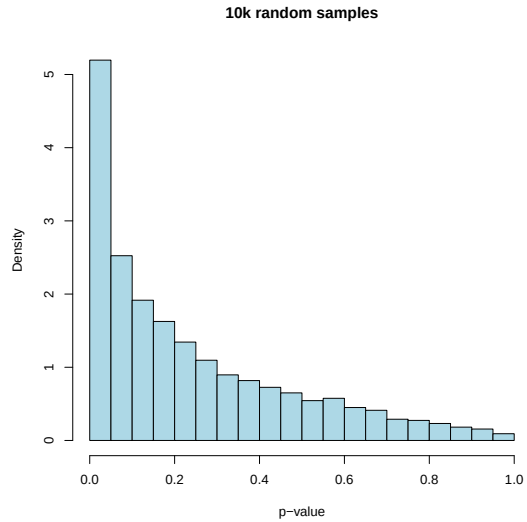


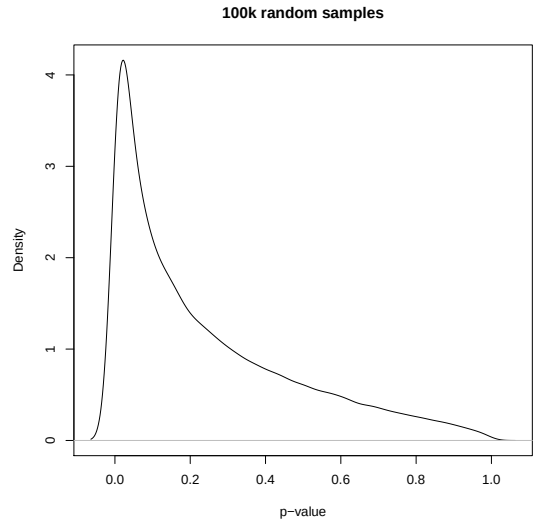


$\bar{x}$=120.9 $\bar{x}$ =118.9 $\cdots$ $\bar{x}$ = 121.2

p-value=0.18 *p*-value=0.86 $\cdots$ *p*-value=0.12







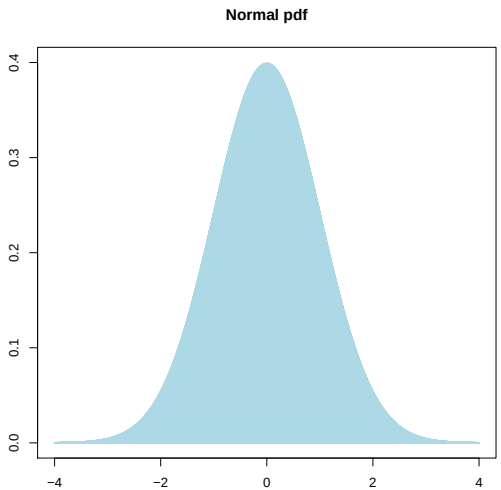
Histogram - and smoothed histogram of *p*-values.

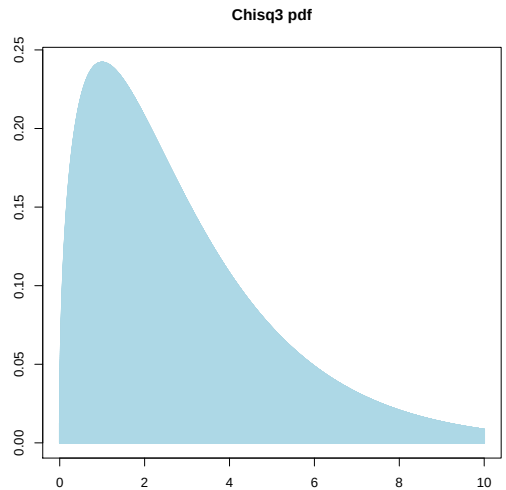
More about the p-value

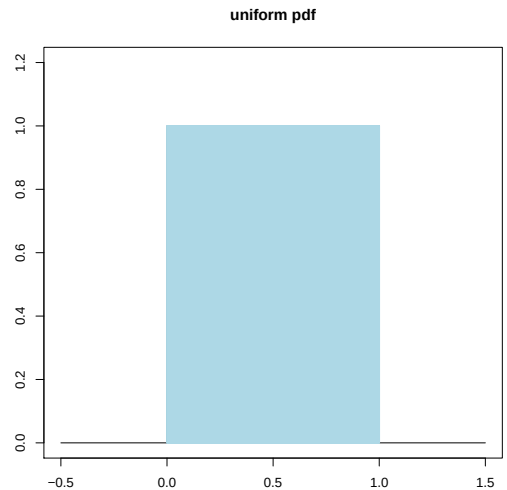
- The p-value is just a function of the random sample and can be regarded as a random variable.
  We had: $P(\bar{X} > \text{observed mean} \mid H_0 \text{ true})$.
- But, isn't the p-value a probability? A number?
- A random variable (like the p-value) has a *probability distribution*.
- What is the distribution of a p-value?

Probability distribution for random variable Y

- Continuous random variable Y (could be the p-value).
- Probability distribution function (pdf): $f(y)$.

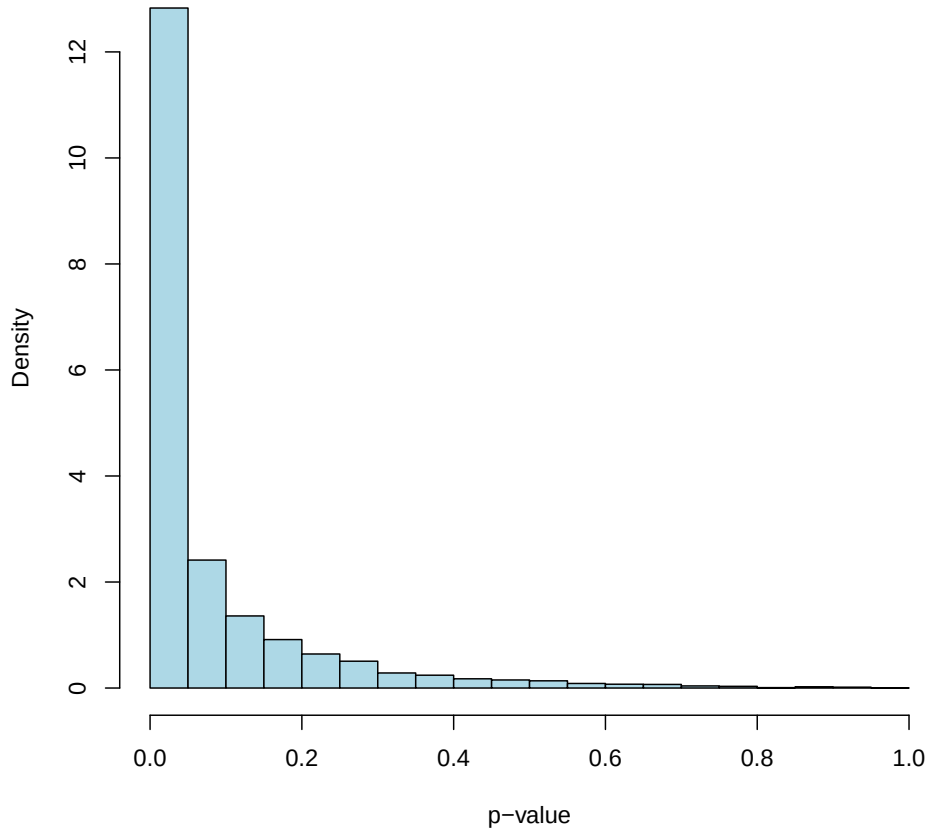






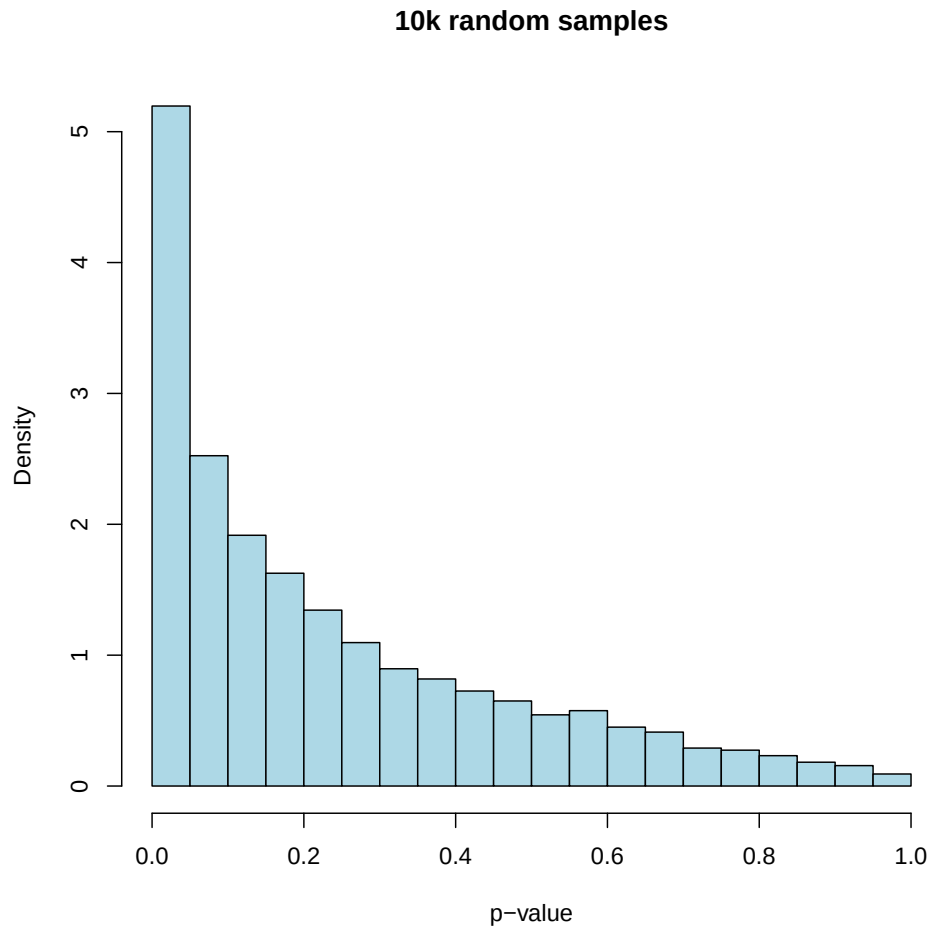
Distribution of p-values for false hypothesis?

Blood pressure example:
Assume that $\mu = 122$ so that H_0 is false, and that we collect a random sample of size 100. What is then the distribution of the p-value?

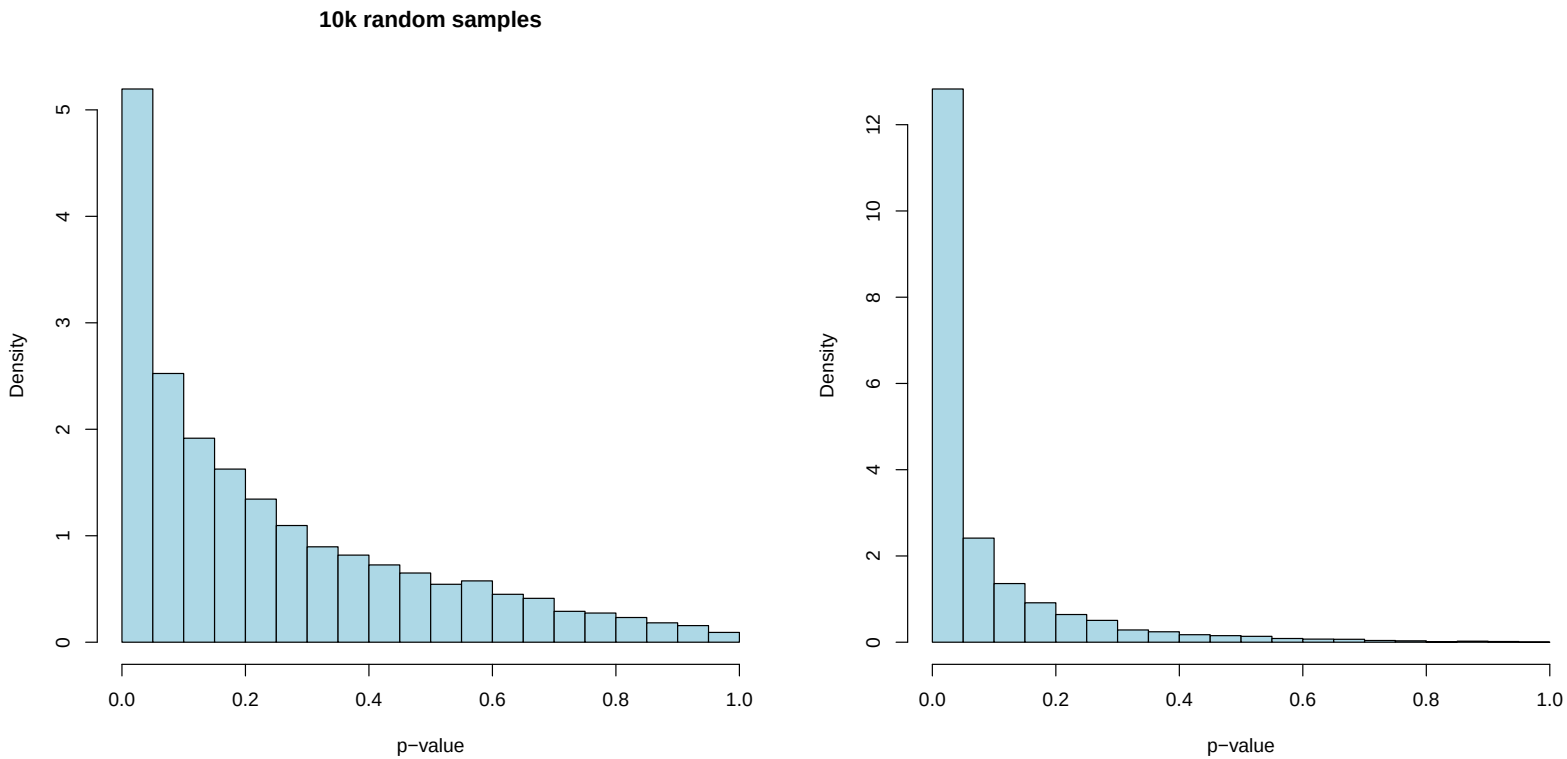


Distribution of p-values for false hypothesis?

Blood pressure example:
Assume that $\mu = 121$ so that H_0 is false, and that we collect a random sample of size 100. What is then the distribution of the p-value?

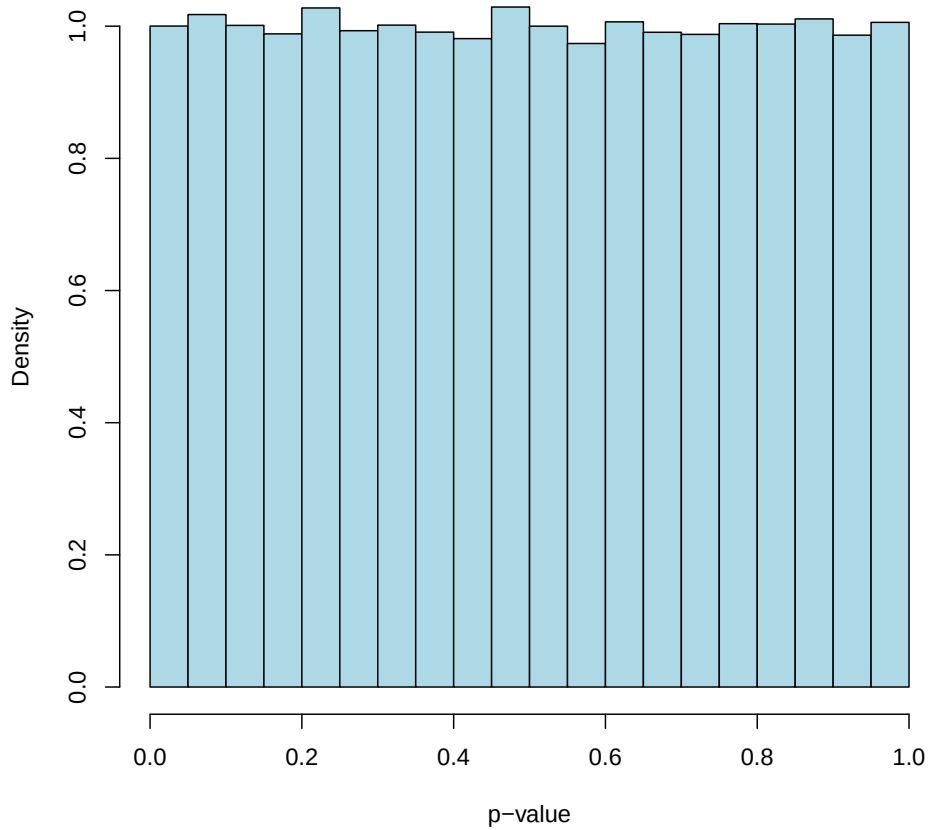


False null: $\mu = 121$ left, and $\mu = 122$ right, when $H_0 : \mu = 120$



Distribution of p-values for true hypothesis?

Blood pressure example:
Assume that $\mu = 120$ so that H_0 is true, and that we collect a random sample of size 100. What is then the distribution of the p-value?



Urban myth: A p-value for a true null hypothesis is close to 1. No, all intervals of equal length are equally probable! =uniform distribution

p-values from true null hypothesis is uniformly distributed

Why is this important:

- so you don't believe the urban myth, and
- it might be useful to understand plots (pdf or cdf) of p-values, and these are often used for quality control of statistical models.

Assume that large values of the test statistic T leads to rejection of the null hypothesis, and that a value t of the test statistic T corresponds to a value w of the p-value W. This means that $P(T \geq t) = P(W \leq w)$. On the other hand the p-value is $P(W \leq w) = P(T \geq t) = w$ when H_0 is true.
This means that $P(W \leq w) = w$ when H_0 is true. If W is a continuous random variable taking values from 0 to 1, the the p-value W must be uniformly distributed over the interval from 0 to 1.
This is true when the p-value is continuous and exact.

Exact *p*-value

If $P(p(\boldsymbol{Y}) \leq \alpha) = \alpha$ for all α, $0 \leq \alpha \leq 1$, the *p*-value is called an *exact* *p*-value.

Valid p-value

A p-value $p(\boldsymbol{Y})$ is *valid* if

$$P(p(\boldsymbol{Y}) \leq \alpha) \leq \alpha$$

for all α, $0 \leq \alpha \leq 1$, whenever H_0 is true, that is, if the p-value is valid, rejection on the basis of the p-value ensures that the probability of type I error does not exceed α.

From single to multiple hypothesis testing

In many situations we are not interested in testing only one hypothesis, but instead m hypotheses.

- In a regression setting m might be the number of covariates in the regression model, and we would test $H_0 : \beta_j = 0$ vs $H_1 : \beta_j \neq 0$ for all $j = 1, \ldots, m$.
- If we have a linear regression with one categorical covariate with k levels, called a one-way analysis of variance model, we might first want to test $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$ against the alternative hypothesis, H_1, that the means of at least two of the k levels are different from each other. If the null hypothesis is rejected we might want to continue to test which of all possible pairs of the means that are different – giving $m = \binom{k}{2}$ hypothesis tests, or compare the mean of all levels to a common reference level μ_1, giving $m = k - 1$ hypothesis tests.

But, can't we still use cut-off α on the p-values to detect significant findings?

Westfall & Young (1993): Multicenter Oat Bran Study

- At each of ten study centers a control vs treated experiment is performed with 20 subjects per group.
- It is common to analyze the data for each center separately, as well as to combine over center.
- T-statistics are computed for each center as

$$\frac{\bar{y}_T - \bar{y}_C}{\sqrt{(s_T^2 + s_C^2)/20}}$$

with p-values calculated as lower tail probabilities from the t-distribution with 38 degrees of freedom.

FIRST Oat Bran Study

Table 1.2 First Multicenter Oat Bran Study Using Simulated Data

Center	Group	$\bar{y}$	s	t-Statistic	p-Value (Lower-Tailed)
1	Treated	219.1	7.0	.30	.616
	Control	218.3	9.8		
2	Treated	212.6	11.3	−1.76	.043*
	Control	218.5	9.8		
3	Treated	207.5	11.6	−1.79	.041*
	Control	213.6	9.9		
4	Treated	212.5	10.4	.76	.774
	Control	209.6	13.5		
5	Treated	211.9	8.5	1.90	.968
	Control	206.6	9.1		
6	Treated	222.3	13.4	.06	.523
	Control	222.1	7.5		
7	Treated	212.0	7.4	.04	.515
	Control	211.9	8.9		
8	Treated	217.4	8.6	.82	.792
	Control	215.0	9.8		
9	Treated	220.7	10.7	1.28	.895
	Control	217.2	6.0		
10	Treated	222.9	9.1	−.45	.326
	Control	224.4	11.6		

* p-value less than .05.

FIRST Oat Bran Study

- Centres 2 and 3 show significant reduction in blood cholestreol for the treatment group.
- Centre 5 happens to show a significant increase, but that is not "noticed" since one-sided tests are performed.
- If the studies were run as uncoordinated trials, it is likely that the two significant studies would be reported and perhaps published in reputable journals.
- The eight nonsignificant studies would go to the file drawer and a "true, confirmed" effect would be established for the two sites where significance is found.
- The centres with insignificant results may decide to collect fresh data, and analyse only the new data.

SECOND Oat Bran Study

Table 1.3 Second Hypothetical Oat Bran Study

Center	Group	$\bar{y}$	s	t-Statistic	p-Value (Lower-Tail
1	Treated	214.6	9.2	1.90	.968
	Control	209.3	8.4		
2	Treated	213.9	8.7	1.21	.884
	Control	210.2	10.5		
3	Treated	217.6	7.6	.59	.720
	Control	216.0	9.5		
4	Treated	215.5	6.2	1.59	.940
	Control	211.7	8.7		
5	Treated	211.6	9.6	1.24	.889
	Control	208.1	8.2		
6	Treated	220.1	8.7	.069	.527
	Control	219.9	9.6		
7	Treated	210.3	5.9	−2.00	.026*
	Control	215.0	8.7		
8	Treated	212.2	9.8	−1.55	.065
	Control	217.7	12.5		
9	Treated	217.3	8.8	.79	.784
	Control	215.0	9.5		
10	Treated	212.2	11.2	.53	.700
	Control	210.5	9.0		

* p-value less than .05.

Oat bran study: lessons to be learned

- These are SIMULATED data with equal means of the control and the treatment group, i.e. the truth is that there are no biological effects of the treatment.
- With simulated data: simple to point to the multiplicity issue as the *cause* for the small p-values for some centres.
- Real studies: not easy to determine if a seen effect is real or not.

Oat bran study: lessons to be learned

- Real studies: not easy to determine if a seen effect is real or not.
- At a particular centre showing significance: scientists would believe that the effect is real, because why should the existence of other centres in the study affect the outcome at the given centre?
- How should one verify that an unusual event is real or artificial?
- The possibility of false positive results is very real, and can lead to serious misinterpretation by analysts: it is human nature to rationalize any dramatic- statistically significant - change.

From single to multiple hypothesis testing

Set-up

- Let us assume that we perform m hypothesis tests,
- giving m p-values and then
- choose a cut-off on the p-values at some value α_{loc} (called a local significance level) to decide if we want to reject each null hypothesis.
- We then reject the null hypotheses where the p-value is smaller than α_{loc}, and this leads to rejection of R hypotheses.

Multiple hypothesis testing set-up

One hypothesis:

	Not reject H_0	Reject H_0
H_0 true	Correct	Type I error
H_0 false	Type II error	Correct

m **hypotheses:**

	Not reject H_0	Reject H_0	Total
H_0 true	U	V	m_0
H_0 false	T	S	$m - m_0$
Total	$m - R$	R	m

- R rejected null hypotheses
- V false positives (type I errors)
- T false negatives (type II errors)

Only m and R are observed. **What should we now control?**

Overall Type I error control (1)

- In some situation one expects that just a few null hypothesis are false,
- therefore a *strict* criterion for controlling an overall version of the Type I error is chosen.
- Family-Wise Error Rate (FWER) is controlled at level α.

  $$\text{FWER} = P(V \geq 1) = P(\text{the number of false positives is } \geq 1)$$

  (remark: V is not observed)
- The FWER can be controlled by defining a *local significance level* α_{LOC} for each test and reject the H_0 of that test if the p-value of the test is less than the α_{LOC}.

Basal metabolic rate and the FTO-gene: revisited

- The gene called FTO is known to be related to obesity
- The basal metabolic rate says how many calories you burn when you rest (hvilemetabolisme).
- Data has been collected for 101 patient from the obesity clinic at St. Olavs Hospital.
- Research question: is there an association between the variant of the FTO gene of the patient and the basal metabolic rate?
- Regression setting, other covariates include age, sex, weight, height, BMI, diet, exercise level, smoking, etc.

If we had not only collected data on this one gene, but instead for many (e.g. $m = 100000$) genetic markers positioned along the chromosome, and then wanted to test m hypotheses, we would not expect to find many true associations. This strategy is called a genome-wide association analysis and for this purpose FWER is usually controlled.

Overall Type I error control for GWA data: FWER control

- GWAS often use $\alpha_{\text{LOC}} = 5 \cdot 10^{-8}$.
- The most popular method controlling the FWER is the Bonferroni method, which can always be used.
- The Bonferroni method might be slightly conservative (too low α_{LOC}), since it is constructed to control FWER for all types of dependency structures between the test statistics for the different hypotheses- including independence.
- `https://arxiv.org/abs/1603.05938`: *Efficient and powerful familywise error control in genome-wide association studies using generalized linear models*, K. K. Halle, Ø. Bakke, S. Djurovic, A. Bye, E. Ryeng, U. Wisløff, O. A. Andreassen, M. Langaas.

Overall Type I error control (2)

- For other types of data one expects that many null hypotheses are false,
- and therefore a less strict criterion for controlling an overall version of the Type I error is chosen.
- The False Discovery Rate (FDR) by Benjamini & Hochberg (1995) is controlled at level α.
- Informally, the FDR is the expected proportion of Type I errors among the rejected hypotheses.

$$\text{FDR} = E(Q) \text{ where by definition}$$

$$Q = \begin{cases} V/R & \text{if } R > 0, \text{ or} \\ 0 & \text{if } R = 0 \end{cases}$$

Hedenfalk et al (2001) gene expression dataset

Available from library(qvalue) from Bioconductor

- The data from the breast cancer gene expression study of Hedenfalk et al. (2001) were obtained and analyzed.
- A comparison was made between 3,226 genes of two mutation types, BRCA1 (7 arrays) and BRCA2 (8 arrays).
- The data included here are p-values, test-statistics, and permutation null test-statistics obtained from a two-sample t-test analysis on a set of 3170 genes, as described in Storey and Tibshirani (2003).

For such gene expression data researchers expect to find may genes that are differently expressed between conditions and therefore the false discovery rate (FDR) is usually controlled.

Hedenfalk I et al. (2001). Gene expression profiles in hereditary breast cancer. *New England Journal of Medicine*, 344: 539-548.
Storey JD and Tibshirani R. (2003). Statistical significance for genome-wide studies. *Proceedings of the National Academy of Sciences*, 100: 9440-9445. http://www.pnas.org/content/100/16/9440.full

Overall Type I error control for gene expression data

- Popular algorithm for controlling the FDR: the Benjamini-Hochberg step-up procedure.
- Focus on minimal interesting biological effect: is possible that you don't want to test *difference between treatments*=0, but instead $\geq$ minimal biological interesting effect.

Multiple testing

- Note from course www-page.
- RecEx5.Problem 2.
- CompulsoryPart3 Problem 2.
- This topic is new on the reading list in 2017.
- It replaces the topics of regularization with the lasso and ridge regression, which will be covered in TMA4268 Statistical Learning.

Summarizing Part 3

with quiz in Kahoot!